Multi-Task Zero-Shot Action Recognition with Prioritised Data Augmentation

Multi-Task Zero-Shot Action Recognition with Prioritised Data Augmentation

Xun Xu, Timothy M. Hospedales and Shaogang Gong

School of EECS, Queen Mary University of London, UK

Abstract. Zero-Shot Learning (ZSL) promises to scale visual recognition by bypassing the conventional model training requirement of annotated examples for every category. This is achieved by establishing a mapping connecting low-level features and a semantic description of the label space, referred as visual-semantic mapping, on auxiliary data. Reusing the learned mapping to project target videos into an embedding space thus allows novel-classes to be recognised by nearest neighbour inference. However, existing ZSL methods suffer from auxiliary-target domain shift intrinsically induced by assuming the same mapping for the disjoint auxiliary and target classes. This compromises the generalisation accuracy of ZSL recognition on the target data. In this work, we improve the ability of ZSL to generalise across this domain shift in both model- and data-centric ways by formulating a visual-semantic mapping with better generalisation properties and a dynamic data re-weighting method to prioritise auxiliary data that are relevant to the target classes. Specifically: (1) We introduce a multi-task visual-semantic mapping to improve generalisation by constraining the semantic mapping parameters to lie on a low-dimensional manifold, (2) We explore prioritised data augmentation by expanding the pool of auxiliary data with additional instances weighted by relevance to the target domain. The proposed new model is applied to the challenging zero-shot action recognition problem to demonstrate its advantages over existing ZSL models.

1 Introduction

Action recognition has long been a central topic in computer vision [1]. A major thrust in action recognition is scaling methods to a wider and finer range of categories [2–4]. The traditional approach to dealing with a growing number of categories is to collect labeled training examples of each new category. This is not scalable, particularly in the case of actions, due to the temporally extended nature of videos compared to images, making annotation (segmentation in *both* space and time) more onerous than for images. In contrast, the Zero-Shot Learning (ZSL) [5, 6] paradigm is gaining significant interest by providing an alternative to classic supervised learning which does not require an ever increasing amount of annotation. Instead of collecting training data for the target categories[1] to be recognised, a classifier is constructed by re-using a visual to

[1] Target and testing all refer to categories (e.g. action classes) to be recognised without labelled examples.

semantic space mapping pre-learned on a training/auxiliary set [2] of totally independent (disjoint) categories. Specifically training class labels are represented in a vector space such as attribute [5, 7] or word-vectors [6, 8]. Such vector representations of class-labels are referred to as *semantic label embeddings* [7]. A mapping (e.g. regression [9] or bilinear model [7]) is learned between low-level visual features and their semantic embeddings. This mapping is assumed to generalise and be re-used to project visual features of target classes into semantic embedding space and matched against target class embeddings.

A fundamental challenge for ZSL is that in the context of supervised learning of the visual-semantic mapping, the ZSL setting violates the traditional assumption of supervised learning [10] – that training and testing data are drawn from the same distribution. Thus its efficacy is reduced by *domain shift* [11–13]. For example, when a regressor is used to map visual features to semantic embedding, the disjoint training and testing classes in ZSL intrinsically require the regressor to generalise out-of-bounds. This inherently limits the accuracy of ZSL recognition. In this work, we address the issue of the generalisation capability of a ZSL mapping regressor from both the model- and data-centric perspectives: (1) by proposing a more robust regression model with better generalisation properties, and (2) improving model learning by augmenting training data with a large additional dataset using dynamic re-weighting to prioritise auxiliary instances and categories that are most relevant (likely to generalise to) the target problem.

Multi-Task Embedding When establishing the mapping between visual features and semantic embeddings, most ZSL methods learn each dimension of this mapping *independently* – whether semantic embedding is discrete as in the case of attributes [5, 7], or continuous as in the case of word vectors [6, 8]. This has the advantage of simplicity, but is more likely to overfit to the training classes and hence suffer from weak generalisation to the disjoint target classes. This is because it treats each dimension of the label in semantic embedding independently despite the labels living on a non-uniform manifold [14] and because many independent mappings result in a large number of parameters to be learned. We denote this conventional approach as Single-Task Learning (STL) due to the independent learning of mappings for each attribute/word dimension. In contrast, we advocate a Multi-Task Learning (MTL) [15, 16, 10] regression approach to mapping visual features and their semantic embeddings. By constraining the mapping parameters of each learning task to lie closely on a low-dimensional manifold, we gain two advantages: (1) Exploiting the relation between the response variables (dimensions of the label embedding), (2) reducing the total number of parameters to fit. The resulting visual-semantic mapping is more robust to the domain shift between ZSL training and testing classes. As a helpful byproduct, the MTL mapping, provides a lower dimensional latent space in which the nearest neighbour (NN) matching required by ZSL can be better performed [17] compared to the usual higher dimensional label semantic embedding space.

[2] Auxiliary and training all refer to categories (e.g. action classes) with labelled data.

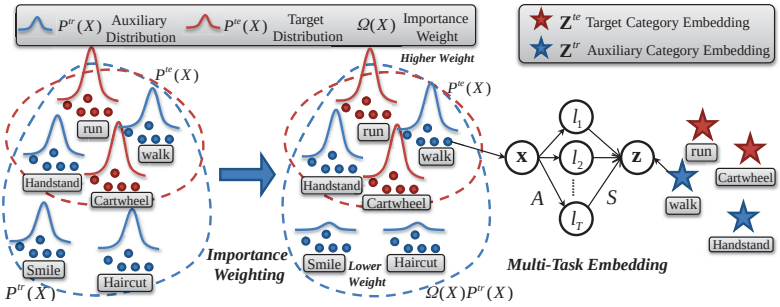

Fig. 1. Two strategies to improve generalisation of visual-semantic mapping in ZSL. Left: Importance weighting to prioritise auxiliary data relevant to the target domain. Right: Learning the mapping from visual features $\mathbf{X}$ to semantic embedding $\mathbf{Z}$ by MTL reduces overfitting, and also provides a latent lower dimensional representation $\{\mathbf{l}_t\}$ to benefit nearest neighbour matching.

Prioritised Auxiliary Data Augmentation for Domain Adaptation
From a data-, rather than model-centric perspective, studies have also attempted to improve the generalisation of ZSL methods by augmenting[3] the auxiliary dataset with additional datasets containing a wider array of classes and instances [9, 18]. The idea is that including a broader additional set should provide better coverage of the visual feature and label embedding spaces, therefore helping to learn a visual-semantic mapping that better generalises to target classes, and thus improves performance when representing and recognising target classes. However, existing studies on exploring this idea have been rather crude, e.g. simply expanding the training dataset by blindly concatenating auxiliary set with additional data [9]. This is not only inefficient but also dangerous, because it does not take into account the (dis)similarity between the extra incorporated data and the target classes for recognition, thus risking *negative transfer* [10]. In this work, we address the issue that auxiliary and target data/categories will have different marginal distributions (Fig 1). We selectively re-weight those relevant instances/classes from the auxiliary data that are expected to improve the the visual-semantic mapping in the context of the specific target classes to be recognised (target domain). We formulate this prioritised data augmentation as a domain adaptation problem by minimizing the discrepancy between the marginal distributions of the auxiliary and target domains. To achieve this, we propose an importance weighting strategy to re-weight each auxiliary instance in order to minimise the discrepancy. Specifically we generalise the classic *Kullback-Leibler Importance Estimation Procedure* (KLIEP) [19, 20] to the zero-shot learning problem – minimising the approximated KL-Divergence between the auxiliary and the target domains, both in terms of visual and semantic label embedding representations.

[3] In this work, data augmentation means exploiting additional data in a wider context from multiple data sources, in contrast to synthesising more artificial variations of one dataset as in deep learning.

2 Related Work

Zero-Shot Learning Zero-shot Learning (ZSL) [5] aims to generalize existing knowledge to recognize new categories without training examples by re-using a mapping learned from visual features to their semantic embeddings. Commonly used label embeddings are semantic attributes [5, 21, 11] and word-vectors [6, 9]. The latter has the advantage of being learned from data without requiring manual annotation. Commonly used visual-semantic mappings include linear [12] and non-linear regression [11, 6, 9], classification [5, 21], and bilinear ranking [7].

Existing ZSL methods suffer from weak generalisation due to the domain-shift induced by disjoint auxiliary-target classes, an issue that has recently been highlighted explicitly in the literature [8, 11–13]. Attempts to address this so far include post-processing heuristics [11–13], sparse coding regularisation [8], and simple blind enlarging of the training set with auxiliary data [9]. In contrast to [8, 9], we focus on: (1) Building a visual-semantic mapping with intrinsically better generalisation properties, and (2) re-weighting the auxiliary set to prioritise auxiliary instances most relevant to the target instances and classes. Our method is complementary to [11, 12] and can benefit from these heuristics.

Zero-Shot Action Recognition Among many ZSL tasks in computer vision, zero-shot action recognition [21, 9, 22–24] is of particular interest because of the lesser availability of *labelled* video compared to image data; and because videos are more difficult to label than static images due to extended temporal duration and more complex ontology. ZSL action recognition is much less studied than still image recognition, and existing video-ZSL methods suffer from the same domain-shift drawbacks highlighted above.

Multi-Task Regression Learning Multi-Task Learning (MTL) [10, 25] aims to improve generalisation in a set of supervised learning tasks by modelling and exploiting shared knowledge across the tasks. Various sharing structures have been proposed to model the relations between tasks. An early study [15] proposed to model the weight vector for each task t as a sum of a shared global task $\mathbf{w}_0$ and task specific parameter vector $\mathbf{w}_t$. However, the assumption of a globally shared underlying task is too strong, and risks inducing *negative transfer* [10]. This motivates the Grouping and Overlapping Multi-Task Learning (GOMTL) [16] framework which instead assumes that each task's weight vector is a task-specific combination of a small set of latent basis tasks. This constrains the parameters of all tasks to lie on a low dimensional manifold.

MTL methods have been studied for action recognition [26–29]. However, all of these studies focus on improving standard *supervised* action recognition with multi-task sharing. For example, considering each of multiple views [28, 29], feature modalities [27], or – most obviously – action categories [26] as different tasks. Multi-view/multi-feature recognition is orthogonal to our work, while the later ones are concerned with supervised recognition, and cannot be generalised to the ZSL scenario. In contrast, we take a very different approach and treat each dimension of the visual-semantic mapping as a task, in order to leverage MTL to improve auxiliary-target generalisation across the disjoint target categories. Fi-

Table 1. Notation Summary

Notation	Description
$n_c^{tr}; n_c^{te}$	Number of training categories ; testing categories
$n_x^{tr}; n_x^{te}$	Number of all training instances; all testing instances
$\mathbf{X} \in \mathbb{R}^{d_x \times n_x}$; $\mathbf{x}_i$	Visual feature matrix for N instances; column representing the i-th instance
$\mathbf{Y} \in \{0,1\}^{n_c \times n_x}$; $\mathbf{y}_i$	Binary class labels for N instances 1-of-n_c encoding; column representing the i-th instance
$\mathbf{V} \in \mathbb{R}^{d_z \times n_c}$;	Semantic label embedding for n_c categories;
$\mathbf{Z} \in \mathbb{R}^{d_z \times n_x}$; $\mathbf{z}_i$	Semantic label embedding for n_x instances; column representing the i-th instance
$\mathbf{W} \in \mathbb{R}^{d_z \times d_x}$; $\mathbf{w}_d$	STL regression coefficient matrix; row representing the regressor for the d-th dimension
$\mathbf{A} \in \mathbb{R}^{T \times d_x}$; $\mathbf{a}_t$	MTL regression coefficient matrix; row representing the regressor for the t-th latent task
$\mathbf{S} \in \mathbb{R}^{d_z \times T}$; $\mathbf{s}_d$	MTL linear combination matrix; row representing linear combination vector for the d-th output
$\mathbf{L} \in \mathbb{R}^{T \times n_x}$; $\mathbf{l}_i$	Latent space embedding for visual instances; column is ith instance
$\omega \in \mathbb{R}^{n_x \times 1}$	weighting vector for auxiliary data
$f : \mathbf{X} \rightarrow \mathbf{Z}$	Visual to semantic mapping function

nally, we note that the use of MTL to learn the visual semantic mapping provides a further benefit of a lower-dimensional space in which zero-shot recognition can be better performed due to being more meaningful for NN matching [17].

Importance Weighting for Domain Adaptation Domain shift is a widely studied problem in transfer learning [10], although it is usually induced by sampling bias [30, 31] or sensor change [32] rather than the disjoint categories in ZSL. Importance weighting (IW) [19, 31] has been one of the main adaptation techniques to address this issue. The idea behind these techniques is to align the auxiliary with the target data to maximise performance on the target data. Importantly the prior work in this area is designed for the standard domain transfer problem in a *supervised* learning setting [33], while we are the first to generalise it to the *zero-shot* learning scenario. The IW technique we generalise is related to another domain adaptation approach based on discovering a feature mapping to minimise the *Maximum Mean Discrepancy* (MMD) [34, 35] between distributions. However MMD, is less appropriate for us due to focus on feature mapping rather than instance reweighing, and our expectation is that only subsets of auxiliary instances will be relevant to the target rather than the holistic auxiliary set.

Contributions This paper contributes both model- and data-centric strategies to improve ZSL action recognition: (1) We formulate learning a more generalisable visual-semantic mapping in ZSL as a multi-task learning problem with a lower-dimensional latent semantic embedding space for more effective matching. (2) We improve visual-semantic regression generalisation by prioritised data augmentation using importance weighting of auxiliary instances relevant to the target domain.

3 Visual-Semantic Mapping with Multi-Task Regression

In ZSL, we aim to recognise action categories $\mathbf{Y}$ given visual features $\mathbf{X}$ where training/auxiliary and testing/target categories do not overlap $\mathcal{Y}^{tr} \cap \mathcal{Y}^{te} = \emptyset$. The key method by which ZSL is achieved is to embed each category label in $\mathcal{Y}$ into a semantic label embedding space $\mathcal{Z}$ which provide a vector representation of any *nameable* category. Table 1 summarises the notation used in the subsequent sections.

3.1 Training a Visual Semantic Mapping

Before presenting our new model, we first introduce briefly the conventional single task learning using regression for visual-semantic mapping[12, 9, 11].

Single-Task Regression Given a matrix $\mathbf{V}$ describing the embedded action names[4], and per-video binary labels $\mathbf{Y}$, we firstly obtain the label embedding of any action label for a video clip as $\mathbf{z}_i = \mathbf{V}\mathbf{y}_i$. We then learn a visual-semantic mapping function $f : \mathcal{X} \rightarrow \mathcal{Z}$ on the training categories. Given a loss function $l(\cdot, \cdot)$, we learn the mapping f by optimising Eq (1) where $\Omega(f)$ denotes regularization on the mapping:

$$\min_f \frac{1}{n_x^{tr}} \sum_{i=1}^{n_x^{tr}} l\left(f(\mathbf{x}_i), \mathbf{z}_i\right) + \Omega(f). \tag{1}$$

The most straightforward choice of mapping f and loss l is linear $f(\mathbf{x}) = \mathbf{W}\mathbf{x}$, and square error respectively, which results in a regularized linear (ridge) regression problem: $l\left(f(\mathbf{x}_i), \mathbf{z}_i\right) = ||\mathbf{z}_i - \mathbf{W}\mathbf{x}_i||_2^2$. A closed-form solution to $\mathbf{W}$ can then be obtained by $\mathbf{W} = \mathbf{Z}\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T + \lambda n_x^{tr}\mathbf{I}\right)^{-1}$. Each row $\mathbf{w}_d$ of regressor $\mathbf{W}$ maps visual feature $\mathbf{x}_i$ to dth dimension of response variable $\mathbf{z}_i$. Since regressors $\{\mathbf{w}_d\}_{d=1\cdots d_z}$ are learned independently from each other this is referred as **single-task learning (STL)** with each $\mathbf{w}_d$ defining one distinct 'task'.

From Single to Multi-Task Regression In the conventional ridge-regression solution to Eq. (1), each task $\mathbf{w}_d$ is effectively learned separately, ignoring any relationship between tasks. We wish to model this relationship by discovering a latent basis of predictors such that tasks $\mathbf{w}_d$ are constructed as linear combinations of T latent tasks $\{\mathbf{a}_t\}_{t=1\cdots T}$. So the dth regression predictor is now modelled as $\mathbf{w}_d = \sum_t s_{dt}\mathbf{a}_t = \mathbf{s}_d^T\mathbf{A}$, where $\mathbf{s}_d$ is the combination coefficient for d-th task. Denoting multi-task regression prediction as $f(\mathbf{x}_i, \mathbf{S}, \mathbf{A})$, we now optimise:

$$\min_{\mathbf{S},\mathbf{A}} \frac{1}{n_x^{tr}} \sum_{i=1}^{n_x^{tr}} l(f(\mathbf{x}_i, \mathbf{S}, \mathbf{A}), \mathbf{z}_i) + \lambda\Omega(S) + \gamma\Psi(\mathbf{A}). \tag{2}$$

Grouping and Overlap Multi-Task Learning An effective method following the MTL design pattern above is GOMTL [16]. GOMTL uses a $\mathbf{W} = \mathbf{S}\mathbf{A}$ task parameter matrix factorisation, where the number of latent tasks T (typically $T < d_z$) is a free parameter. Requiring the combination coefficients $\mathbf{s}_t$ to be sparse, via a ℓ_1 regulariser, the loss is written as

$$\min_{\{\mathbf{s}_t\},\mathbf{A}} \sum_{t=1}^{T} \frac{1}{n_x^{tr}} \sum_{i=1}^{n_x^{tr}} ||\mathbf{z}_{t,i} - \mathbf{s}_t\mathbf{A}\mathbf{x}_i|| + \lambda \sum_{t=1}^{T} ||\mathbf{s}_t||_1 + \gamma||\mathbf{A}||_F^2 \tag{3}$$

This can be solved by iteratively updating $\mathbf{A}$ and $\mathbf{S}$. When $\mathbf{A}$ is fixed, loss function reduces to a standard L1 regularized (LASSO) regression problem

[4] To deal with multi-word compound action category names, e.g. "Apply Eye Makeup", we apply a simple average, summing the component word vectors [9, 11].

that can be efficiently solved by Alternating Direction Method of Multipliers (ADMM) [36]. When $\mathbf{S}$ is fixed, we take the derivative of loss w.r.t. $\mathbf{A}$ and obtain the following gradient, allowing efficient solution by gradient descent:

$$\nabla_{\mathbf{A}} = \tfrac{2}{n_x^{tr}} \sum_{t=1}^{T} (-\mathbf{s}_t \mathbf{X}\mathbf{Z}^T + \mathbf{s}_t^T \mathbf{X}\mathbf{X}^T \mathbf{s}_t \mathbf{A}) + 2\lambda_A \mathbf{A} \tag{4}$$

Regularized Multi-Task Learning (RMTL) The classic RMTL method [15] models task parameters as the sum of a globally shared and task specific parameter vector: $\mathbf{w}_t = \mathbf{a}_0 + \mathbf{a}_t$. It can be seen that this corresponds to a special case of GOMTL's $\mathbf{W} = \mathbf{S}\mathbf{A}$ predictor matrix factorisation [25]. Here there are $T = d_z + 1$ latent tasks, a fixed task combination vector $\mathbf{s}_t = [1 \quad \mathbf{1}(t = 1) \quad \mathbf{1}(t = 2) \cdots \mathbf{1}(t = d_z)]^T$ where $\mathbf{1}(\cdot)$ is the indicator function and $A = \left[\mathbf{a}_0^T \mathbf{a}_1^T \cdots \mathbf{a}_{d_z}^T\right]^T$.

Explicit Multi-Task Embedding (MTE) In GOMTL Eq (3), it can be seen that the label embedding $\mathbf{z}_i$ is approximated from the data by the mapping $\mathbf{s}_t \mathbf{A}\mathbf{x}_i$, and this approximation is reached by combination via the latent representation $\mathbf{A}\mathbf{x}_i$. While GOMTL defines this space implicitly via the learned $\mathbf{A}$, we propose to model it explicitly as $\mathbf{l}_i \approx \mathbf{A}\mathbf{x}_i$. This is so the actual projections $\mathbf{l}_i$ in this latent space can be regularised explicitly, in order to learn a latent space which generalises better to test data, and hence improves ZSL matching later.

Specifically, we split the GOMTL loss $||\mathbf{z}_i - \mathbf{S}\mathbf{A}\mathbf{x}_i||_2^2$ into two parts: $||\mathbf{l}_i - \mathbf{A}\mathbf{x}_i||_2^2$ and $||\mathbf{z}_i - \mathbf{S}\mathbf{l}_i||_2^2$ to learn the mapping to the latent space, and from the latent space to the label embedding respectively. This allows us to place additional regularization on $\mathbf{l}_i$ to avoid extreme values in the latent space and thus later improve neighbour matching (Section 3.2). Given the large and high dimensional video datasets, we apply Frobenius norm on $\mathbf{S}$ in contrast to GOMTL's ℓ_1.

$$\begin{aligned} \min_{\{\mathbf{s}_t\},\mathbf{A},\{\mathbf{l}_i\}} \quad & \sum_{t=1}^{T} \frac{1}{n_x^{tr}} \sum_{i=1}^{n_x^{tr}} \left(||\mathbf{z}_{t,i} - \mathbf{s}_t \mathbf{l}_i||_2^2 + ||\mathbf{l}_i - \mathbf{A}\mathbf{x}_i||_2^2 \right) + \\ & \lambda_S \sum_{t=1}^{T} ||\mathbf{s}_t||_2^2 + \lambda_A ||\mathbf{A}||_F^2 + \lambda_L \sum_{i=1}^{n_x^{tr}} ||\mathbf{l}_i||_2^2 \end{aligned} \tag{5}$$

Our explicit multi-task embedding has similarities to [18], but our purpose is multi-task regression for ZSL, rather than embedding for video descriptions. To solve our explicit embedding model we iteratively solve $\mathbf{L}$,$\mathbf{A}$ and $\mathbf{S}$ while fixing the other two. With the ℓ_2 norm on $\mathbf{S}$, this has a convenient closed-form solution to each parameter:

$$\begin{aligned} \mathbf{L} &= (\mathbf{S}^T\mathbf{S} + (\lambda_L n_x^{tr} + 1)\mathbf{I})^{-1}(\mathbf{S}^T\mathbf{Z} + \mathbf{A}\mathbf{X}) \\ \mathbf{S} &= \mathbf{Z}\mathbf{L}^T(\mathbf{L}\mathbf{L}^T + \lambda_S n_x^{tr}\mathbf{I})^{-1} \\ \mathbf{A} &= \mathbf{L}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda_A n_x^{tr}\mathbf{I})^{-1} \end{aligned} \tag{6}$$

3.2 Zero-Shot Action Recognition

We consider two alternative NN matching methods for zero-shot action prediction that use the MTL mappings described above.

Distributed Space Matching Given a trained visual-semantic regression f, we project testing set visual feature $\mathbf{x}^{te}$ into the semantic label embedding space. The standard strategy [9, 11, 12] is then to employ NN matching in this space for zero-shot recognition. Specifically, given the matrix of label embeddings for each target category name $\mathbf{V}^{te}$, and using cosine distance norm, the testing video $\mathbf{x}^{te}$ are classified by:

$$\mathbf{y}^* = arg\min_{\mathbf{y}^*} ||\mathbf{V}^{te}\mathbf{y}^* - f(\mathbf{x}^{te})|| \tag{7}$$

where $f(\mathbf{x}^{te}) = \mathbf{W}\mathbf{x}^{te}$ for STL and $f(\mathbf{x}^{te}) = \mathbf{S}\mathbf{A}\mathbf{x}^{te}$ for MTL.

Latent Space Matching MTL methods provide an alternative to matching in label space: Matching in the latent space. The representation of testing data in this space is the output of latent regressors $\mathbf{l}_{te} = \mathbf{A}\mathbf{x}^{te}$ (Eq. (5)). To get the representation of testing categories in the latent space we invert the combination matrix $\mathbf{S}$ to project target category names $\mathbf{V}^{te}$ into latent space. Specifically we classify by Eq. (8), where $(\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T$ is the Moore-Penrose pseudoinverse.

$$\mathbf{y}^* = arg\min_{\mathbf{y}^*} ||(\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\mathbf{V}^{te}\mathbf{y}^* - \mathbf{A}\mathbf{X}^{te}|| \tag{8}$$

NN matching in the latent space is better than in semantic label space because: (i) the dimension is lower $T < d_z$, and (ii) we have explicitly regularised the latent space to be well behaved (Eq. (5)).

4 Importance Weighting

An advantage of ZSL with word-vector based label embedding is that the visual-semantic mapping can potentially be improved by augmenting auxiliary data with additional examples from other datasets[9], providing more data to learn a better visual-semantic mapping. However, simply aggregating auxiliary and additional datasets is not ideal as including irrelevant data risks 'negative transfer'. Therefore we are motivated to develop methodology to prioritise augmented auxiliary data that is useful for a particular ZSL recognition scenario. Specifically, we learn a per-instance weighting $\omega(\mathbf{x})$ on the auxiliary dataset $\mathbf{X}^{tr}$ to adjust each instance's contribution according to relevance to the target domain. Because Importance Weighting (IW) adapts auxiliary data to the target domain, we assume a transductive setting with access to testing data $\mathbf{X}^{te}$.

Kullback-Leibler Importance Estimation Procedure (KLIEP) We first introduce the way to estimate a per-instance auxiliary-data weight given the distribution of target data $\mathbf{X}^{te}$. This is based on the idea [19] of minimizing the KL-divergence (D_{KL}) between training $p^{tr}(\mathbf{x})$ and testing data distribution $p^{te}(\mathbf{x})$ via learning a weighting function $\omega(\mathbf{x})$. This is formalised in Eq. (9):

$$\begin{aligned} &\min_{\omega} D_{KL}(p^{te}(\mathbf{x})|\omega(\mathbf{x})p^{tr}(\mathbf{x})) = \int p^{te}(\mathbf{x}) \log \frac{p^{te}(\mathbf{x})}{\omega(\mathbf{x})p^{tr}(\mathbf{x})} d\mathbf{x} \\ &\min_{\omega} \int p^{te}(\mathbf{x}) \log \frac{p^{te}(\mathbf{x})}{p^{tr}(\mathbf{x})} d\mathbf{x} - \int p^{te}(\mathbf{x}) \log \omega(\mathbf{x}) d\mathbf{x} \end{aligned} \tag{9}$$

The first term is fixed w.r.t. $\omega(\mathbf{x})$ so the objective to optimise is:

$$\min_{\omega} - \int p^{te}(\mathbf{x}) \log \omega(\mathbf{x}) d\mathbf{x} \approx -\frac{1}{n_x^{te}} \sum_{i=1}^{n_x^{te}} \log \omega(\mathbf{x}_i) \tag{10}$$

Aligning Both Visual Features and Labels KLIEP is conventionally used for domain adaptation by reweighting instances [19, 33]. In the case of transductive ZSL, we have the target data $\mathbf{X}^{te}$ and category labels $\mathbf{Z}^{te}$ respectively, although not instance-label association which is to be predicted. In this case we can further improve ZSL by extending KLIEP to align training and testing sets in both visual feature and category sense[5]. Specifically, we minimise the kullback-leibler divergence between the target and auxiliary in terms of both the visual and category distributions:

$$\begin{aligned}
&\min_{\omega_x,\omega_z} D_{KL}(p^{te}(X)||\omega_x(\mathbf{X})p^{tr}(\mathbf{X})) + D_{KL}(p^{te}(\mathbf{Z})||\omega_z(\mathbf{Z})p^{tr}(\mathbf{Z})) \\
&\min_{\omega_x,\omega_z} \int p^{te}(\mathbf{X}) \log \frac{p^{te}(\mathbf{X})}{\omega_x(\mathbf{X})p^{tr}(\mathbf{X})} d\mathbf{X} + \int p^{te}(\mathbf{Z}) \log \frac{p^{te}(\mathbf{Z})}{\omega_z(\mathbf{Z})p^{tr}(\mathbf{Z})} d\mathbf{Z} \\
&\min_{\omega_x,\omega_z} -\frac{1}{n_x^{te}} \sum \log \omega_x(\mathbf{x}_i^{te}) - \frac{1}{n_x^{te}} \sum \log \omega_z(\mathbf{z}_i^{te})
\end{aligned} \tag{11}$$

Given both $\mathbf{X}^{te}$ and $\mathbf{Z}^{te}$, we construct the weighting functions as a combination of Gaussian kernels centered at the testing data and categories in Eq (12). Here $\omega(\mathbf{x}, \mathbf{z})$ extends the previous notation $\omega(\mathbf{x})$ to indicate giving a weight to each training instance given visual feature $\mathbf{x}$ and class name embedding $\mathbf{z}$. So if there are n_x^{tr} instances, $\omega(\mathbf{x}, \mathbf{z})$ returns a weight vector of length n_x^{tr}.

$$\begin{aligned}
&\omega(\mathbf{x}, \mathbf{z}) = \omega_x(\mathbf{x}) + \omega_z(\mathbf{z}) \quad \omega_x(\mathbf{x}) = \sum_{i=1}^{n_x^{te}} \alpha_i \phi(\mathbf{x}, \mathbf{x}_i^{te}), \\
&\omega_z(\mathbf{z}) = \sum_{i=1}^{n_x^{te}} \beta_j \phi(\mathbf{z}, \mathbf{z}_i^{te}) \quad \phi(\mathbf{x}, \mathbf{x}_i^{te}) = exp\left(-\frac{||\mathbf{x} - \mathbf{x}_i^{te}||^2}{2\sigma^2}\right)
\end{aligned} \tag{12}$$

For ease of formulation, we denote $\mathbf{a} = [\alpha_1 \cdots \alpha_{n_x^{te}}]^T$, $\mathbf{b} = [\beta_1 \cdots \beta_{n_x^{te}}]^T$, $\Phi_{\mathbf{a}}(\mathbf{x}) = [\phi(\mathbf{x}, \mathbf{x}_1^{te}) \cdots \phi(\mathbf{x}, \mathbf{x}_{n_x^{te}}^{te})]^T$ and $\Phi_{\mathbf{b}}(\mathbf{z}) = [\phi(\mathbf{z}, \mathbf{z}_1^{te}) \cdots \phi(\mathbf{z}, \mathbf{z}_{n_x^{te}}^{te})]^T$. The optimization can be thus written as

$$\min_{\mathbf{a},\mathbf{b}} -\frac{1}{n_x^{te}} \sum_{i=1}^{n_x^{te}} \log \mathbf{a}^T \Phi_{\mathbf{a}}(\mathbf{x}_i^{te}) - \frac{1}{n_x^{te}} \sum_{i=1}^{n_x^{te}} \log \mathbf{b}^T \Phi_{\mathbf{b}}(\mathbf{z}_i^{te}), \quad s.t. \quad \frac{1}{n_x^{tr}} \sum_{i=1}^{n_x^{tr}} \omega(\mathbf{x}_i^{tr}, \mathbf{z}_i^{tr}) = 1 \tag{13}$$

The above constrained optimization problem is convex w.r.t. both $\mathbf{a}$ and $\mathbf{b}$. It can be solved by interior point methods using the derivatives in Eq. (14):

[5] KLEIP with labels was studied by [20], but they assumed the target joint distribution of $\mathbf{X}$ and $\mathbf{Z}$ is known. So [20] is only suitable for traditional supervised learning with labeled target examples of $\mathbf{z}_i$ and $\mathbf{x}_i$ in correspondence. In our case we have the videos to classify and the zero-shot category names, but the assignment of names to videos is our task rather than prior knowledge.

$$\nabla \mathbf{a} = -\frac{1}{n_x^{te}} \sum_{i=1}^{n_x^{te}} \frac{1}{\mathbf{a}^T \Phi_{\mathbf{a}}(\mathbf{x}_i^{te})} \Phi_{\mathbf{a}}(\mathbf{x}_i^{te}), \quad \nabla \mathbf{b} = -\frac{1}{n_x^{te}} \sum_{i=1}^{n_x^{te}} \frac{1}{\mathbf{b}^T \Phi_{\mathbf{b}}(\mathbf{z}_i^{te})} \Phi_{\mathbf{b}}(\mathbf{z}_i^{te}) \tag{14}$$

Weighted Visual-Semantic Regression Given per-instance weights ω estimated above, we can rewrite the loss function for both single-task ridge regression and multi-task regression in Sec 3.1 as $\omega_i l(f(\mathbf{x}_i, \mathbf{A}), \mathbf{z}_i)$ and $\omega_i l(f(\mathbf{x}_i, \mathbf{S}, \mathbf{A}), \mathbf{z}_i)$ respectively. All our loss functions have quadratic form, so the weight can be expressed inside the quadratic loss e.g. $\omega_i ||\mathbf{z}_i - \mathbf{W}\mathbf{x}_i||_2^2 = ||\mathbf{z}_i\sqrt{\omega_i} - \mathbf{W}\mathbf{x}_i\sqrt{\omega_i}||_2^2$. Thus to incorporate the weight information we simply replace the original semantic embedding matrix with $\tilde{\mathbf{z}}_i = \mathbf{z}_i\sqrt{\omega_i}$ and data matrix with $\tilde{\mathbf{x}}_i = \mathbf{x}_i\sqrt{\omega_i}$.

5 Experiments

Datasets and Settings We evaluated our contributions on three human action recognition datasets, HMDB51 [3], UCF101 [4] and Olympic Sports [37]. They contain 6766, 13320, 783 videos and 51, 101, 16 categories respectively. For all datasets we extract improved trajectory feature (ITF) [38], a state-of-the-art space-time feature representation for action recognition. We use Fisher Vectors (FV) [39] to encode three raw descriptors (HOG, HOF and MBH). Each descriptor is reduced to half of its original dimension by PCA, resulting in a 198 dim representation. Then we randomly sample 256,000 descriptors from all videos and learn a Gaussian Mixture with 128 components to obtain the FVs. The final dimension of FV encoded feature is $2 \times 128 \times 198 = 50688$ dimensions. For the label-embedding, we use 300-dimensional word2vec [40]. We use $T = n_c^{tr}$ latent tasks, and cross-validation to determine regularisation strength hyper-parameters for the models[6].

5.1 Visual-semantic Mappings for Zero-Shot Action Recognition

Evaluation Criteria To evaluate zero-shot action recognition, we divide each dataset evenly into training and testing parts with 5 random splits. Using classification accuracy for HMDB51 and UCF101 and average precision for Olympic Sports as the evaluation metric, the average and standard deviation over the 5 splits are reported for each dataset.

Compared Methods We study the efficacy of our contributions by evaluating the different visual-semantic mappings presented in Sec 3.1. We compare MTL-regression methods with conventional STL Ridge Regression (denoted **RR**) for ZSL. For RR/STL, nearest neighbour matching is used to recognise target categories. Note that the RR+NN method here corresponds to the core strategy used by [9, 11, 12]. The multi-task models we explore include: **RMTL** [15]: assumes each task's predictor is the sum of a global latent vector and a task-specific vector. **GOMTL** [16]: Uses a predictor-matrix factorisation assumption in which

[6] Ridge Regression (RR) has 15M (300×50688) parameters, whilst for HMDB51 where T=25, GOMTL and MTE have 1.27M ($50688 \times 25 + 25 \times 300$) parameters.

Table 2. Visual-semantic mappings for zero-shot action recognition: MTL (✓) versus STL (✗). Latent matching (✓) versus distributed (✗) matching

ZSL Model	MTL	Latent Matching	HMDB51	UCF101	Olympic Sports
RR	✗	NA	18.3 ± 2.1	14.5 ± 0.9	40.9 ± 10.1
RMTL [15]	✓	✗	18.5 ± 2.1	14.6 ± 1.1	41.1 ± 10.0
RMTL [15]	✓	✓	18.7 ± 1.7	14.7 ± 1.0	41.1 ± 10.0
GOMTL [16]	✓	✗	18.5 ± 2.2	13.1 ± 1.5	43.5 ± 8.8
GOMTL [16]	✓	✓	18.9 ± 1.0	14.9 ± 1.5	44.5 ± 8.5
MTE	✓	✗	18.7 ± 2.2	14.2 ± 1.3	$\mathbf{44.5 \pm 8.2}$
MTE	✓	✓	$\mathbf{19.7 \pm 1.6}$	$\mathbf{15.8 \pm 1.3}$	44.3 ± 8.1

tasks' predictors lie on a low-dimensional subspace. **Multi-Task Embedding (MTE)**: Our model differs from GOMTL in that it explicitly models and regularises a lower dimensional latent space. For the multi-task methods, we also compare the ZSL matching strategies introduced in Section 3.2: **Distributed:** Standard NN matching (Eq. (7)), and **Latent:** our proposed latent-space matching (Eq. (8)).

Results: The comparison of single task ridge regression with our multi-task methods is presented in Table 2. From these results we make the following observations: (i) Overall our multi-task methods improve on the corresponding single-task baseline of RR. MTL regression (RMTL, GOMTL and MTE) improves single-task ridge regression by $5-10\%$ in relative terms, with the biggest margins visible on the Olympic Sports dataset. (ii) Within multi-task models, the GOMTL with sparse ℓ_1 regularization outperforms RMTL. This suggests learning the task combination $\mathbf{S}$ from data is better than fixing it as in RMTL. (iii) Our MTE generally outperforms other multi-task methods supporting the explicit modelling and regularisation of the latent space. (iv) In most cases, NN matching in the latent space improve zero-shot performance. This is likely due to the lower dimension of the latent space compared to the dimension of the original word vector embedding, making NN matching more meaningful [17].

5.2 Importance Weighted Data Augmentation

We next evaluate the impact of importance weighting in data augmentation for zero-shot action recognition. We perform the same 5 random split benchmark for each dataset. For data augmentation, we augment each dataset's training split with the data from all other datasets. For instance, for ZSL on HMDB51 we augment the training data with all videos from UCF101 and Olympic Sports.

Compared Methods We study the impact of the data augmentation methods: **Naive DA:** Naive Data Augmentation [9, 41] simply assigns equal weight to each auxiliary training sample. **Visual KLIEP:** The auxiliary data is aligned with the testing sample distribution $\mathbf{X}^{te}$ (Eq. (9)). **Category KLIEP:** The auxiliary categories are aligned with testing category distribution $\mathbf{Z}^{te}$. This is achieved by the same prodcedure in Eq. (9) by replacing $\mathbf{x}$ with $\mathbf{z}$. **Full KLIEP:** The distribution of both samples $\mathbf{X}^{te}$ and categories $\mathbf{Z}^{te}$ is used to reweight the auxiliary data (Eq. (13)).

Table 3. Data augmentation and importance weighting for ZSL action recognition.

ZSL Model	Weighting Model	HMDB51	UCF101	OlympicSports
RR	Naive DA	21.9 ± 2.4	19.4 ± 1.7	46.5 ± 9.4
MTE	Naive DA	$\mathbf{23.4 \pm 3.4}$	$\mathbf{20.9 \pm 1.5}$	$\mathbf{49.4 \pm 8.8}$
RR	Visual KLIEP	23.2 ± 2.7	20.3 ± 1.6	47.2 ± 9.3
RR	Category KLIEP	23.0 ± 2.1	20.2 ± 1.6	51.8 ± 8.7
RR	Full KLIEP	23.7 ± 2.7	20.7 ± 1.4	51.3 ± 9.0
MTE	Visual KLIEP	23.4 ± 2.8	20.8 ± 2.0	51.4 ± 9.2
MTE	Category KLIEP	23.3 ± 2.4	20.9 ± 1.7	50.9 ± 8.3
MTE	Full KLIEP	$\mathbf{23.9 \pm 3.0}$	$\mathbf{21.9 \pm 2.7}$	$\mathbf{52.3 \pm 8.1}$

Results: From the results in Table 3, we draw the conclusions: (i) Both the baseline single task learning (STL) method and our Multi-Task Embedding (MTE) improve with Naive DA (compare unaugmented results in Table 2), (ii) The Visual, Category, and Full visual+category-based weightings all improve on Naive DA in the case of STL RR. (iii) We see that our MTE with Full KLIEP augmentation performs the best overall. The ability of KLIEP to improve on Naive DA suggests that the auxiliary data is indeed of variable relevance to the target data, and selectively re-weighing the auxiliary data is important. (iv) For KLIEP-based DA, either Visual or Category DA provides most of the improvement, with relatively less improvement obtained by using both together.

Alternative Models We also compare against previous state-of-the-art methods including those driven by both attributes and word-vector category embeddings. **DAP/IAP** [5]: Direct/Indirect attribute prediction are classic attribute-based zero-shot recognition models based on training SVM classifiers independently for each attribute, and using a probabilistic model to match attribute predictions with target classes. **HAA**: We implement a simplified version of the Human Actions by Attributes model [21]: We first train attribute detection SVMs, and test samples are assigned to categories based on cosine distance between their vector of attribute predictions and the target classes' attribute vectors. **SVE** [9]: Support vector regression was adopted to learn the visual to semantic mapping. **ESZSL** [42]: Embarrassingly Simple Zero-Shot Learning defines the loss function as the mean square error on label prediction in contrast to the regression loss defined in other baseline models. **SJE**: Structured Joint Embedding [7] employed a triplet hinge loss. The objective is to enforce relevant labels having higher projection values from visual features than those of non-relevant labels. **UDA**: The Unsupervised Domain Adaptation model [22] learns dictionary on auxiliary data and adapts it to the target data as a constraint on the target dictionary rather than blindly using the same dictionary. This work combines both attribute and word vector embeddings.

Comparison Versus State of the Art: Table 4 compares our models with various contemporary and state-of-the-art models. For clear comparison, we indicate for each method which embedding ((**W**)ordvector / (**A**)ttribute) and feature (our FV, or BoW) are used, as well as whether it has a transductive dependency on the test data (**TD**) or exploits additional augmenting data (**Aug**). From these results we conclude that: (i) Although data augmentation has a big

Table 4. Comparison versus state of the art. Embed: Label embedding, Feat: Visual feature used, Aug: Data augmentation required? TD: Transductive Requirement?

Method	Embed	Feat	TD	Aug	HMDB51	UCF101	Olympic Sports
MTE	W	FV	X	X	19.7 ± 1.6	15.8 ± 1.3	44.3 ± 8.1
MTE + Full KLIEP	W	FV	✓	✓	23.9 ± 3.0	21.9 ± 2.7	52.3 ± 8.1
MTE + Full KLIEP + PP	W	FV	✓	✓	$\mathbf{24.8 \pm 2.2}$	$\mathbf{22.9 \pm 3.3}$	$\mathbf{56.6 \pm 7.7}$
MTE	A	FV	X	X	N/A	18.3 ± 1.7	55.6 ± 11.3
DAP [5] - CVPR 2009	A	FV	X	X	N/A	15.9 ± 1.2	45.4 ± 12.8
IAP [5] - CVPR 2009	A	FV	X	X	N/A	16.7 ± 1.1	42.3 ± 12.5
HAA [21] - CVPR 2011	A	FV	X	X	N/A	14.9 ± 0.8	46.1 ± 12.4
SVE [9] - ICIP 2015	W	BoW	X	X	14.9 ± 1.8	12.0 ± 1.4	N/A
SVE [9] - ICIP 2015	W	BoW	✓	X	15.6 ± 0.7	16.5 ± 2.4	N/A
SVE [9] - ICIP 2015	W	BoW	X	✓	19.3 ± 4.0	13.1 ± 2.0	N/A
SVE [9] - ICIP 2015	W	BoW	✓	✓	22.8 ± 2.6	18.4 ± 1.4	N/A
ESZSL [42] - ICML 2015	W	FV	X	X	18.5 ± 2.0	15.0 ± 1.3	39.6 ± 9.6
ESZSL [42] - ICML 2015	W	FV	X	✓	22.7 ± 3.5	18.7 ± 1.6	51.4 ± 8.3
ESZSL [42] - ICML 2015	A	FV	X	X	N/A	17.1 ± 1.2	53.9 ± 10.8
SJE [7] - CVPR 2015	W	FV	X	X	13.3 ± 2.4	9.9 ± 1.4	28.6 ± 4.9
SJE [7] - CVPR 2015	A	FV	X	X	N/A	12.0 ± 1.2	47.5 ± 14.8
UDA [22] - ICCV 2015	A	FV	✓	X	N/A	13.2 ± 1.9	N/A
UDA [22] - ICCV 2015	A+W	FV	✓	X	N/A	14.0 ± 1.8	N/A

impact, our non-transductive and no data augmentation method (MTE) generally outperforms prior alternatives due to learning an effective latent matching space robust to the train/test class shift; (ii) The performance of our MTE with word-vector embedding is strong when compared with DAP/IAP/HAA/ESZSL even with attribute embedding. Given the same attribute embedding, MTE outperforms all state-of-the-art models due to the discovery of latent attributes from the original attribute space; (iii) Moreover, given importance weighting on auxiliary data, our method (MTE + Full KLIEP) with word-vector embedding performs the best overall – including against [9] which also exploits data augmentation; (iv) Finally, our method is synergistic to the post processing self-training approach [11] as well as the hubness strategies [12], which further explains the advantages of our approach (MTE + Full KLIEP + PP) over other methods.

5.3 Qualitative Results and Further Analysis

Importance Weighting: To visualise the impact of our IW, we randomly select 4 / 16 classes as target / auxiliary sets respectively. We then estimate the weight on the 16 auxiliary video classes according to the Full KLIEP (Section 4). Examples of the auxiliary video weightings are presented in Fig 2. We observe that auxiliary classes semantically related to the targets are given higher weight e.g. HandstandPushups→Cartwheel in first sample, SalsaSpin→Hug and Sword Exercise → Fencing in the second sample. While the visually and semantically less relevant auxiliary videos are given much lower weights.

Multi-task Embedding: We next qualitatively illustrate single versus multi-task visual-semantic mappings. Specifically we take 5 classes to be recognized and visualise their data after visual-semantic projection by tSNE [43]. A comparison between the representations generated by single-task (RR) and multi-task (MTE) mappings is given in Fig 3. The multi-task embedding discovers data in a

Fig. 2. Visualisation of Full KLIEP auxiliary data weighting. Left: 4 target videos with category names. Right: 16 auxiliary videos with bars indicating the estimated weights.

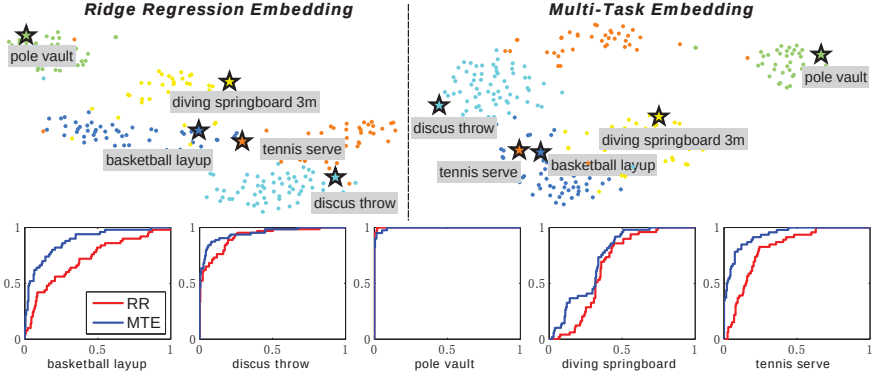

Fig. 3. Qualitative comparison between single-task ridge regression (RR) and multi-task embedding (MTE). The tSNE plots show the testing data and categories labels in the semantic embedding space. ROC curves are evaluated for each target category.

lower dimension latent space where NN classification becomes more meaningful. The improved representation is illustrated by computing the ROC curve for each target category, as seen in Fig 3. MTE provides improved detection over RR, demonstrating the better generalisation of this representation.

6 Conclusion

In this work, we focused on zero-shot action recognition from the perspective of improving generalisation of the visual-semantic mapping across the disjoint train/test class gap. We propose both model- and data-centric improvements to a traditional regression-based pipeline by respectively, multi-task embedding – to minimise overfit of the train data and to build a lower dimensional latent matching space; and prioritising data augmentation by importance weighting – to best exploit auxiliary data for the recognition of target categories. Our experiments on a set of contemporary action-recognition benchmarks demonstrate the impact of both our contributions and show state-of-the-art results overall.

References

1. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: A review. ACM Computing Surveys **43**(3) (2011)
2. Schüldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local SVM approach. In: ICPR. (2004)
3. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: A large video database for human motion recognition. In: ICCV. (2011)
4. Soomro, K., Zamir, A., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
5. Lampert, C., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR. (2009)
6. Socher, R., Ganjoo, M.: Zero-shot learning through cross-modal transfer. In: NIPS. (2013)
7. Akata, Z., Reed, S., Walter, D., Lee, H., Schiele, B.: Evaluation of Output Embeddings for Fine-Grained Image Classification. In: CVPR. (2015)
8. Fu, Z., Xiang, T., Kodirov, E., Gong, S.: Zero-Shot Object Recognition by Semantic Manifold Distance. In: CVPR. (2015)
9. Xu, X., Hospedales, T., Gong, S.: Semantic embedding space for zero-shot action recognition. In: ICIP. (2015)
10. Pan, S.J., Yang, Q.: A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering **22**(10) (2010)
11. Fu, Y., Hospedales, T.M., Xiang, T., Gong, S.: Transductive Multi-view Zero-Shot Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **37** (2015)
12. Dinu, G., Lazaridou, A., Baroni, M.: Improving zero-shot learning by mitigating the hubness problem. In: ICLR. (2015)
13. Lazaridou, A., Dinu, G., Baroni, M.: Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In: Proceedings of ACL, Association for Computational Linguistics (2015)
14. Mahadevan, S., Chandar, S.: Reasoning about linguistic regularities in word embeddings using matrix manifolds. arXiv preprint (2015)
15. Evgeniou, T., Pontil, M.: Regularized multi-task learning. In: ACM SIGKDD. (2004)
16. Kumar, A., Daum, H., Iii, H.D.: Learning Task Grouping and Overlap in Multi-task Learning. In: ICML. (2012)
17. Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is nearest neighbor meaningful? Database Theory (1999)
18. Habibian, A., Mensink, T., Snoek, C.G.M.: VideoStory : A New Multimedia Embedding for Few-Example Recognition and Translation of Events. In: ACM MultiMedia. (2014)
19. Sugiyama, M., Nakajima, S., Kashima, H., Von Bünau, P., Kawanabe, M.: Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation. In: NIPS. (2007)
20. Garcke, J., Vanck, T.: Importance Weighted Inductive Transfer Learning for Regression. In: ECMLPKDD. (2014)
21. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: CVPR. (2011)
22. Kodirov, E., Xiang, T., Fu, Z., Gong, S.: Unsupervised Domain Adaptation for Zero-Shot Learning. In: ICCV. (2015)

23. Gan, C., Yang, Y., Zhu, L., Zhao, D., Zhuang, Y.: Recognizing an action using its name: A knowledge-based approach. International Journal of Computer Vision (2016)
24. Chang, X., Yang, Y., Long, G., Zhang, C., Hauptmann, A.G.: Dynamic concept composition for zero-example event detection. In: AAAI. (2016)
25. Yang, Y., Hospedales, T.M.: A Unified Perspective on Multi-domain and Multi-task Learning. ICLR (2015)
26. Zhou, Q., Wang, G., Jia, K., Zhao, Q.: Learning to share latent tasks for action recognition. In: ICCV. (2013)
27. Yuan, C., Hu, W., Tian, G., Yang, S., Wang, H.: Multi-task sparse learning with beta process prior for action recognition. In: CVPR. (2013)
28. Liu, A.A., Xu, N., Su, Y.T., Lin, H., Hao, T., Yang, Z.X.: Single/multi-view human action recognition via regularized multi-task learning. Neurocomputing (2015)
29. Mahasseni, B., Todorovic, S.: Latent multitask learning for view-invariant action recognition. In: ICCV. (2013)
30. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR. (2011)
31. Huang, J., Gretton, A., Borgwardt, K.M., Schölkopf, B., Smola, A.J.: Correcting Sample Selection Bias by Unlabeled Data. In: NIPS. (2007)
32. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: ECCV. (2010)
33. Pardoe, D., Stone, P.: Boosting for Regression Transfer. In: ICML. (2010)
34. Gretton, A., Borgwardt, K.M., Rasch, M., Schölkopf, B., Smola, A.J.: A kernel method for the two-sample-problem. In: NIPS. (2006)
35. Baktashmotlagh, M., Harandi, M., Lovell, B., Salzmann, M.: Unsupervised domain adaptation by domain invariant projection. In: ICCV. (2013)
36. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine Learning **3**(1) (2011)
37. Niebles, J.C., Chen, C.W., Fei-Fei, L.: Modeling temporal structure of decomposable motion segments for activity classification. In: ECCV. (2010)
38. Wang, H., Oneata, D., Verbeek, J., Schmid, C., Wang, H., Oneata, D., Verbeek, J., A, C.S.: A robust and efficient video representation for action recognition. International Journal of Computer Vision (2015)
39. Perronnin, F., Sánchez, J., Mensink, T.: Improving the Fisher kernel for large-scale image classification. In: ECCV. (2010)
40. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. In: NIPS. (2013)
41. Xu, X., Hospedales, T., Gong, S.: Zero-shot action recognition by word-vector embedding. arXiv preprint arXiv:1511.04458 (2015)
42. Romera-paredes, B., Torr, P.H.S.: An embarrassingly simple approach to zero-shot learning. ICML (2015)
43. Van Der Maaten, L.: Accelerating t-sne using tree-based algorithms. The Journal of Machine Learning Research (2014)